Handling User Interruptions in an Embodied Conversational Agent

Nigel Crook
Oxford University Computing Laboratory
Wolfson Building, Parks Rd
Oxford, United Kingdom
nigc@comlab.ox.ac.uk

Cameron Smith
School of Computing
University of Teesside
Middlesbrough, United Kingdom
c.g.smith@tees.ac.uk

Marc Cavazza
School of Computing
University of Teesside
Middlesbrough, United Kingdom
m.o.cavazza@tees.ac.uk

Stephen Pulman
Oxford University Computing Laboratory
Wolfson Building, Parks Rd
Oxford, United Kingdom
sgp@clg.ox.ac.uk

Roger Moore
Department of Computer Science
University of Sheffield, Regent Court
211 Portobello Street,
Sheffield, United Kingdom
r.k.moore@dcs.shef.ac.uk

Johan Boye
School of Computer Science and Communications
KTH Royal Institute of Technology
100 44 Stockholm, Sweden
johan.boye@speechact.se

ABSTRACT

We present a mechanism for handling "barge-in" interruptions from a user who is engaged in a 'social' conversation with an Embodied Conversational Agent (ECA). The ECA is designed to recognise and be empathetic to the emotional state of the user. Occasionally, the ECA will attempt to positively influence the user's emotional state through the conversation. A characteristic of these conversations is that both the user and the ECA will at times speak long, multi sentence utterances as the conversation progresses. The generation of long utterances from the ECA creates opportunities for the user to barge-in whilst the ECA is speaking. Furthermore, the long ECA utterances may even *provoke* a user interruption since they often include advice to the user about how they should deal with difficult or stressful situations that have arisen. This paper outlines an approach to handling user barge-in interruptions in conversations with an ECA and describes its implementation in the Companions English demonstrator.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design

Keywords

barge-in interruptions, dialogue management, affective strategies, embodied conversational agents

1. INTRODUCTION

As ECAs become progressively more realistic in their expressions and behaviours, it is becoming increasingly important that they should also be able to handle phenomena that occur frequently in inter human conversations, such as multi-party dialogue [14] and interruptions [9]. Most current spoken language dialogue (SLD) systems operate on a strict turn-by-turn basis, with the user and system alternating. In natural dialogues, however, not only do conversationalists often interrupt one another, but they also have strategies to avoid being interrupted ('holding the turn') and if they are interrupted, are able to rapidly re-plan their dialogue intentions. Many SLD systems are described as supporting one kind of interruption, known as 'barge-in' [9, 12, 8], but this usually refers to an ability of the system to stop issuing a prompt if the user starts to reply to it before it is finished. In realistic dialogues, there is more to the handling of interruptions than this. In the current paper, we describe a particular approach to interruption handling in an affective dialogue system system aimed at 'social' conversation, that is, conversation whose purpose is to develop and maintain a supportive relationship with the user rather than to achieve some tightly-defined task such as booking an air flight, hiring a car, or paying a bill.

The prototype ECA presented here has been developed as part of the Companions project, which aims at developing an ECA that supports natural conversation as opposed to task-specific dialogue (Figure 1). The system supports conversation with the user about her working day in the office, and will be referred to as the "How Was Your Day?" (henceforth HYWD) prototype [6]. In this system the ECA provides a sympathetic hearing to the user's difficulties, as well as relevant encouragement or advice. One of our postulates for realistic conversation was to depart from the strict turn-taking of task-oriented dialogue and allow both the user and the system to produce lengthy utterances. To support this, user input is analysed using Information Extraction techniques to recognise the entities mentioned in the input

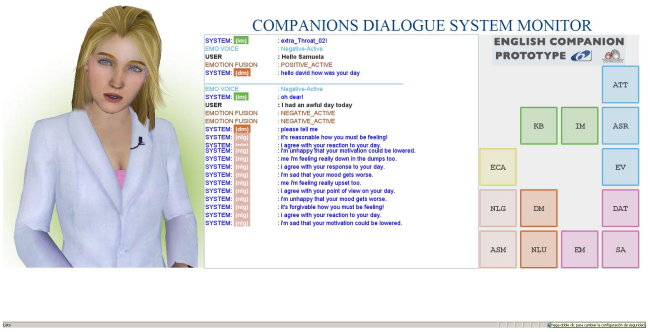

Figure 1: The HWYD prototype ECA interface.

and the main grammatical relationships they take part in. These may be found on general syntactic grounds (subject, object, etc.) or by application specific patterns (e.g. 'argument between X and Y'). The system also attempts to determine referents for pronouns and definite descriptions that may occur in long user utterances. The agent uses planning techniques to generate multi-sentence system utterances offering appropriate advice depending on the information extracted from the long user utterances. Emotional input to the ECA is based on emotional speech recognition and sentiment analysis of the speech recognition transcript, both of which are fused in an emotional model driving the ECA responses. The information extracted from the user's utterance is combined with the fused emotional analysis form the system's appraisal of the user's situation. Since the ECA will at various stages issue advice to the user in the form of long utterances (around 50 words) it is highly likely that these may be interrupted by the user in the course of a realistic conversation.

2. PREVIOUS WORK ON "BARGE-IN" INTERRUPTIONS

This paper presents a method for handling so called 'barge-in' interruptions from the user during a 'social dialogue' with an ECA. A 'barge-in' interruption occurs when the user's begins to speak whilst the ECA is speaking, so that the user's interrupting utterance overlaps with the ECA's speech. A primary task in handling barge-in interruptions is detecting when a genuine user interruption has taken place and differentiating these from other vocal events in the user's speech (e.g. backchannel). Previous work in this area falls into two broad approaches [11]: ones that analyse the acoustic signals from the user's speech to detect features such as prolonged intensity or voicing which indicate a sustained vocal output from the user [1], and ones that use ASR based methods such as language models for barge-in detection. [12, 9, 8]. The method of barge-in detection presented in this paper is based on an analysis of the acoustic signals of the user's speech, and so falls into the first of these groups. Much of the previous work on handling barge-in interruptions has been done in the context of task oriented dialogues such as telephony-based spoken language systems [10, 2]. In contrast, the method presented here is in the context of a 'social dialogue' between the user and an ECA. A characteristic of these dialogues is that both the user and the ECA will speak long, multi sentence utterances as the conversation

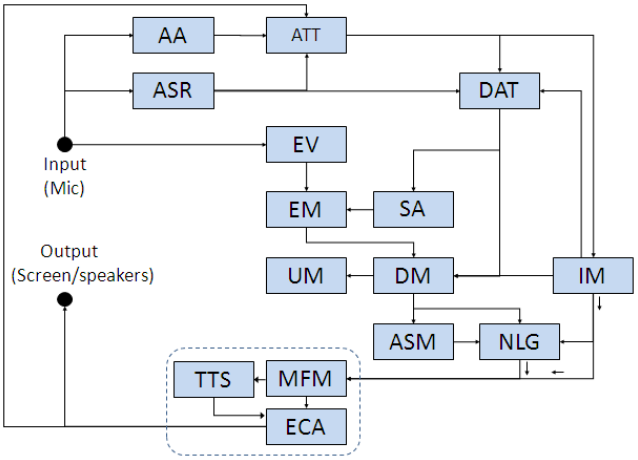

Figure 2: A schematic diagram of the system architecture.

progresses. The user has greater opportunity, therefore, to barge-in whilst the ECA is speaking. The particular challenge for the ECA in this context is to decide how to respond to this interruption. The approach to interruption handling implemented the HWYD prototype and presented in this paper was inspired by cognitive architectures such as Rodney Brooks' subsumption architecture [4, 5].

3. THE COMPANIONS PROJECT DEMONSTRATOR

3.1 System Architecture

The HWYD prototype is composed of 15 modules that communicate through a loosely coupled multi-hub architecture (Inamode) that supports message passing (Figure 2). The user was fitted with a close-talking noise-cancelling bluetooth microphone which maintained a good separation between his voice and the TTS output. Signals from the user's speech are processed by the Automatic Speech Recogniser (ASR - Nuance's Dragon NaturallySpeaking), the Acoustic Analysis module (AA), which extracts low level features from the acoustic signal, and EmoVoice (EV), which performs real-time recognition of the user's emotions from the acoustic properties of their speech. The Acoustic Turn-Taking module (ATT) takes these acoustic features and makes a decision as to whether the user has started or stopped speaking. The Dialogue Act Tagger (DAT) stores the text output of the ASR module until it receives a signal from ATT signifying that the user has stopped speaking. The DAT then segments the text and labels each segment with a dialogue act. These labeled segments are then processed by the Sentiment Analysis (SA) module, which tags each for positive, negative or neutral valance. The Emotional Model (EM) fuses these valance tagged segments of text with the emotional classification of the corresponding speech signal generated by EmoVoice. This is then passed on to the Dialogue Manager (DM) which parses the text segments to extract information from the user's utterances. The DM identifies the major topics of the user's utterance and tags them with the fused emotional label that it received from the EM. This information is then passed on to the Affective

Strategy Module.

3.2 Affective Strategy Module

The HWYD prototype seeks to provide a conversational tone to the dialogue with the user by both attempting to understand lengthy, narrative utterances from the user and replying in kind. These replies require both an understanding of the topics discussed and an appreciation of the emotional state of the user. It is the role of the Affective Strategy Module (ASM) to take the information gathered for both of these aspects and to generate an appropriate narrative response commonly in the form a long, multi sentence utterance to be spoken by the ECA.

The goal of the ASM is to positively influence the user's emotional state through this narrative response. This involves selecting an appropriate affective strategy based on the user's situation and applying this to the generation of a plan of communication acts. This plan is then used to generate the individual utterances (which fit together to compose the larger narrative utterance) which the ECA will use in its response to the user. There are a range of affective strategies available which cover various situations and are selected based on the effect the situation will have on the user and the user's response to that situation. Typically this might consist of a user expressing worry about a potential event which results in the selection of a reassuring strategy that attempts to downplay the threat.

Our approach is based on Bremond's narrative theory of influence [3] in which a character's expectation of a given outcome can be used as a basis for influencing that character. For example, a character anticipating a loss is more readily influenced by discussion of how that loss can be averted or reduced. Bremond's theory comprises two aspects: firstly, the proposal of an ontology of influencing roles (in terms of prototypical narrative situations) and secondly, the relation of influence to communicative acts such as warnings or reassurance. We have applied this approach to our conversations with the user using a default structure for events so as to better allow the identification of influencing factors.

The first stage in the production of the plan of communication acts is to carry out an appraisal of the information received from the conversation with the user. This includes a full analysis of the main topic of discussion. The ASM looks at the effect it will have on the user (in terms of either an improvement or a deterioration in the user's situation) and the likelihood of this outcome occurring. Additional topics of discussion are then reviewed to determine if they will have any influence on the main topic and whether this will be a negative or a positive influence for the user. The emotional state of the user is also examined to determine whether the user's reaction to the situation is appropriate and whether the user's anticipation of the outcome matches the Companion's anticipation. This information is then used to select an appropriate affective strategy.

The final stage in the production of the plan is to use the information gathered from the appraisal, along with the selected affective strategy, to select appropriate communicative acts within the plan. The ASM looks at three main areas when constructing a plan: the first deals with the user's emotional state and provides an appropriate response to acknowledge this, the second provides comments on the main topic based on the selected affective strategy and the third provides a summary to reinforce the chosen affective strategy. The second area can be divided into further areas commenting on different aspects of the user's situation: comments on the appropriateness of the user's emotional reaction, comments on the appropriateness of the user's anticipated outcome, comments on how the user should respond to various influencing factors as well as comments on how the user should respond to the anticipated outcome. The ASM employs a Hierarchical Task Network planner with a heuristic-based selection process [7, 13] to decompose each area, allowing the appraisal information to determine the resulting plan while still affording variability in the plan's composition.

4. INTERRUPTION MANAGEMENT

The generation of long system utterances by the ASM creates opportunities for the user to interrupt the system whilst the ECA is speaking. Furthermore, these long ECA utterances may even *provoke* a user interruption since they often advise the user about how to deal with difficult or stressful situations that they encounter. Here is an example of a long ECA utterance that provokes an angry interruption from the user:

> **ECA:** You should try to calm down. You are doing the right thing about the technical problems. Unfortunately you are correct about them. You can't handle the problems on your own. You should not worry about missing the meeting. Also *USER INTERRUPTS* try to not *ECA DISONTINUES PLANNED UTTERANCE*
>
> **User:** Not worry about missing the meeting! Do you know how important that meeting was!

Example 1: A long system utterance that is interrupted by the user.

The HWYD prototype is based on a pipeline model and so it has no central controller. The handling of a user interrupt therefore involves the cooperative effort of several modules in the system. This is achieved through a staged message passing process. Interruptions are thus handled in the following seven stages:

- Stage 1: ECA outputs a long utterance
- Stage 2: User interrupts ECA
- Stage 3: IM pauses ECA output
- Stage 4: MFM informs DM of where the ECA's utterance was stopped
- Stage 5: ATT signals the end of utterance
- Stage 6: IM and DM respond to the interruption
- Stage 7: ASM continues, replans or aborts the interrupted utterance

The Acoustic Turn Taking (ATT) module is informed whenever the ECA starts or stops speaking (Stage 1). If the user begins to speak while the ECA is still speaking, the ATT makes a decision as to whether this was a "genuine" interruption or simply backchannel or other noise. An interrupt is detected when the intensity of the acoustic signal is such that the user is believed to be talking at the same time

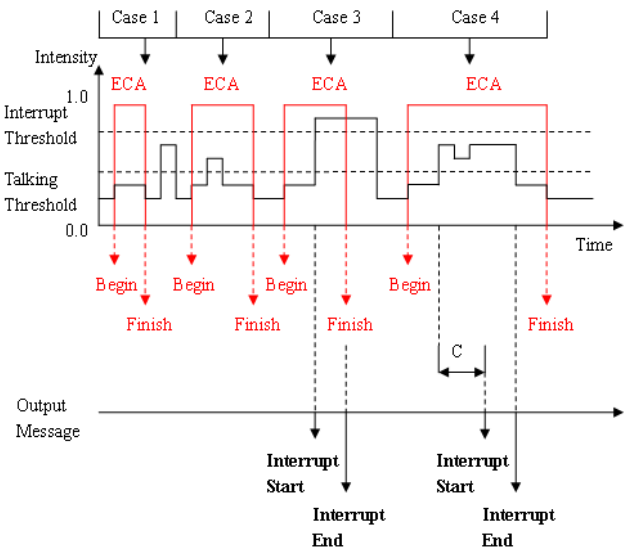

Figure 3: Interrupt detection.

as the ECA, and has talked for more than a certain duration or above a set intensity threshold (the "interrupt" threshold). These conditions aim to avoid treating back-channel acknowledgments of what the avatar has said as an interruption. Figure 3 shows the conditions under which interruption detection occurs. The Talking Threshold is used to determine whether the user is speaking whilst the Interrupt Threshold is used to determine whether the user is speaking at a high enough intensity to be considered an interrupt. The interrupt duration C is used to determine whether the user has talked for long enough whilst the ECA is speaking for this to be considered an interrupt. Note that either of the conditions 'high enough' or 'long enough' will trigger the recognition of an interrupt. The four cases shown in Figure 3 correspond to the following situations in the dialogue:

Case 1: ECA talks, then there is a pause, then user talks (i.e. normal scenario) - no interrupt detected.

Case 2: ECA talks, user starts talking while ECA is talking. But user says something brief and not very loud - user is assumed to be acknowledging ECA, so no interrupt detected.

Case 3: ECA talks, user starts talking while ECA is talking and this is loud (whether brief or not) - user may be shouting, assume this is an interrupt.

Case 4: ECA talks, user starts talking while ECA is talking and this is not brief (longer than interrupt duration constant C) - user is interrupting.

If the ATT concludes that a genuine user interruption has taken place, then it informs the Interruption Manager (IM) module of this (Stage 2). Note that the processing of the user's interrupt by the acoustic modules continues as normal at this point. When the IM receives notification of the interrupt, it first sends a request to the Multi-Modal Fission Module (MFM) to pause the output of ECA speech (Stage 3) and to give the ECA a look of surprise or irritation at being interrupted. Then the IM broadcasts a notification of the user interrupt to all modules. This is done so that modules are alerted to the fact that the previous system turn was not complete and to be aware that the next user utterance they process is an interrupting utterance.

One of the key issues in dealing with a user interruption is knowing what prompted the user to interrupt. In Example 1 it is clear that the user is reacting to the ECA's statement "You should not worry about missing the meeting". Of course, this may not always be the case. The user could be interrupting for all sorts of reasons (e.g. they need to get the bus, go to the toilet, etc). However, in this system it is assumed that it was the preceding ECA utterance that is the most likely cause of the user's interruption. So in Stage 4 the MFM informs the DM where the interruption occurred in the list of utterances planned by the ASM, indicating how much of the long system utterance the ECA managed to say before stopping in response to the user interruption.

When the user stops speaking, the ATT informs the IM that the user's interrupting utterance has ended (Stage 5). The IM now begins to track the processing of the user's interrupting utterance through the various modules of the system. This is done using a System State Model (SSM - Figure 4) which is a two-level Finite State Machine. On the lower level, the SSM models the state of each module in terms of whether it is idle (Id), processing a message (Pr) or waiting for an incoming messages to complete its processing (Pe). On the upper level, the SSM tracks whether each agent in the conversation (system and user) is speaking (Sp), silent (Si), or in the case of the user, interrupting (In). This tracking of the processing of the user's interrupting utterance is necessary to ensure that the ECA responds to the interruption within a realistic time frame. A small set of heuristics that are conditioned on the current state of the SSM and on the length of time since the end of the user's interrupting utterance are used to define this realistic time frame. If these heuristics deem that the processing of the user's interrupting utterance is taking too long, the IM will instruct the DM to output a randomly selected generic acknowledgment of the interruption from the ECA (e.g. "I'm sorry").

Once the processing of the user's interrupting utterance has reached the DM, it needs to decide how to respond to that interruption. We have begun analysing recorded and transcribed conversations between people in order to characterize the way humans typically respond to interruptions. This analysis is still at an early stage and could not inform the interruption response mechanism encoded in the version of the HWYD prototype presented here. Pending the outcome of this investigation, we propose three types of responses to user interrupts: continue, replan and abort.

The DM would choose to continue the ECA's interrupted utterance if the user's utterance does not provide any new information. For example, if the interrupting utterance in Example 1 was "I couldn't agree with you more", then it would be reasonable for the DM to decide to continue the ECA's planned utterances from the point where the interruption took place.

The DM would choose to replan the ECA's utterance if the user's utterance does indeed provide new information.

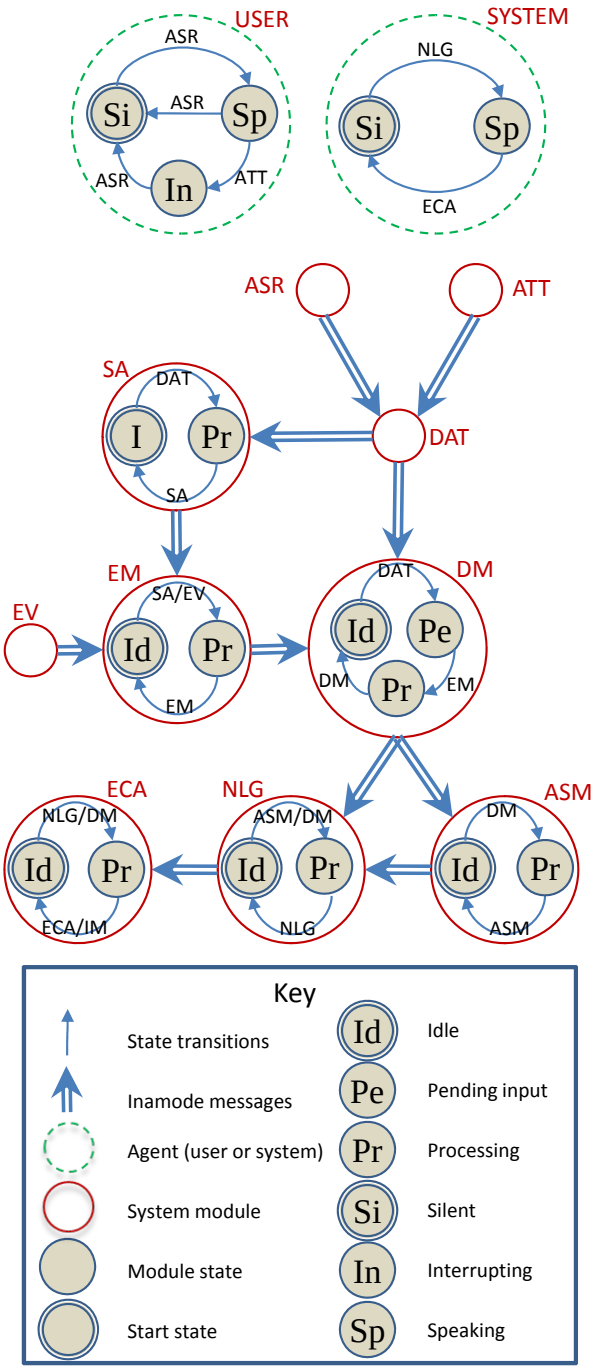

Figure 4: The System State Model which monitors the processing of the user's interrupting utterance.

This would be the case if the user's interrupting utterance in Example 1 corrected what the system had just said with something like "I'm not worried at all about missing the meeting". The replan is necessary because the current plan generated by the ASM was generated from a set of assumptions which have now been shown to be false or incomplete.

The DM would choose to abort the ECA's utterance if the user's utterance rejects the ECA's current dialogue strategy. An abort would be necessary if the user's interrupting utterance was something like "Don't talk to me about about work, I'm not in the mood".

Differentiating between these three interrupt situations is very challenging (see section 6). The current operational version of the HWYD prototype provides a limited analysis of the situation, generally favouring the abort option. Work continues on the DM in building mechanisms that can analyse the user's interrupting utterance to better make this distinction.

Whenever the ASM receives a replan signal from the DM it uses the new information from the DM to regenerate the plan for the utterance. A full re-appraisal of the information from the Dialogue Manager is carried out to ensure no false assumptions are retained from the previous plan. A new plan is then generated using the chosen affective strategy and other information from this re-appraisal. However, as part of the previous plan has already been relayed to the user (and thus invoked the interruption), it may not be appropriate to relay all of an entirely new plan. To do so would risk both repetition and overwhelming the user rather than creating a balanced conversation. Instead the ASM uses the placement of the user's interruption (with respect to the overall length of the previous plan) to determine the length of the new plan. As such, an interruption at the start of the previous plan will result in a full plan being generated while an interruption at the end will result in a much smaller, more focused regenerated plan.

5. EXAMPLE INTERRUPTIONS AND ECA RESPONSES

The following three samples are fragments taken from dialogues with the HWYD prototype. Each shows an instance of a user interruption with either a 'continue', 'replan' or 'abort' response from the ECA.

User: Good morning Samuela (POSITIVE-ACTIVE)

ECA: hello david how was your day

User: My day has been a bit dramatic (POSITIVE-ACTIVE)

ECA: I'm sorry to hear that! please tell me

User: I have just learnt that my department is going to be merged with the Barcelona one (POSITIVE-ACTIVE)

ECA: Oh! really?! you seem a bit too ecstatic! i'm doubtful that your *USER INTERRUPTS* workload could be maintained. *ECA DISONTINUES PLANNED UTTERANCE*

User interruption: I don't want to move to the other office (NEGATIVE-ACTIVE)

ECA: I'm sorry to hear that! no reason to be so aggravated! it's right to be down about such a bad situation.

Example 2: A user interruption that leads to a 'replan' response from the ECA

User: I was late because of the traffic (NEUTRAL)

ECA: how disappointing!

User: And I missed the meeting (NEGATIVE-ACTIVE)

ECA: me i'm feeling really unhappy myself. missing the meeting is perhaps a bad thing! it's right to be unhappy about such *USER INTERRUPTS* a bad situation. *ECA DISONTINUES PLANNED UTTERANCE*

User interrupting: And my boss is really upset (NEGATIVE-ACTIVE)

ECA: what was the meeting about

Example 3: A user interruption with a 'continue' response from the ECA

User: hi Sam (NEGATIVE-ACTIVE)

ECA: hello David how was your day

User: I had a dreadful day to day we had a meeting this morning where they said we might be merged with the Barcelona office (NEGATIVE-PASSIVE)

ECA: I realise things are not great *USER INTERRUPTS*

User interrupting: you telling me they're not great is really upsetting (NEUTRAL)

ECA: what was the meeting about

User: it was about to be a project I'm working on at the moment

Example 4: A user interruption with an 'abort' response from the ECA

6. CONCLUSION

This paper has described the mechanism that was implemented in the HWYD prototype for handling barge-in interruptions from the user during conversations with an ECA. The distinctive contribution of this work in comparison to previous work on interruption handling is that 'barge-in' interruptions are handled in the context of 'social' conversations rather than task-oriented dialogues. The handling of user interrupts in the social conversational setting has particular challenges. One is concerned with the primary aim of the conversation which is for the ECA to maintain and enhance a supporting relationship with the user. User barge-in interrupts in this context can be a strong indicator that the ECA's attempt to maintain a supporting relationship with the user is failing (as demonstrated in Example 1). It is important, therefore, that some care is taken over how the ECA's responds to such interruptions.

We have proposed three responses to barge-in interruptions: 'continue', 'abort' and 'replan'. Another challenge of handling user barge-in interruptions in this context is in discerning when to continue, abort or replan an interrupted ECA utterance. This is a difficult problem because the differences in user utterances which might lead to, say, a continue response being favoured over a replan or abort response might be quite subtle. Consider, for example, the case where the user's interrupting utterance was something like "I couldn't agree with you more". This is not backchannel as such and the most appropriate response is to it would be for the ECA to continue with it's original utterance as planned. However, the interruption "I couldn't agree with you less" is an expression of the user's disagreement with the ECA which would most likely lead to a replanning or an abort of the original ECA utterance. Consequently, a successful strategy for differentiating between user interrupting utterances that require continue, abort and replan responses will require more than a shallow interpretation of the interrupting utterance. Furthermore, the interrupting utterance must to be considered in the context of the ECA utterance that provoked the interruption. We are currently working to expand the DM's ability to analyse the user's interrupting utterance to enable it to differentiate between continue, abort and replan responses to user barge-in interruptions.

7. ACKNOWLEDGMENTS

This work was partially funded by the Companions project (http://www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

8. ADDITIONAL AUTHORS

9. REFERENCES

[1] B. Balentine and D. P. Morgan. *How to build speech recogniton applications - A style guide for telephony dialogs*. Enterprise Integration Group, San Ramon, CA, 1999.

[2] J. Barnett and M. Singh. Designing a portable spoken dialogue system. In E. Maier, M. Mast, and S. LuperFoy, editors, *ECAI Workshop on Dialogue Processing in Spoken Language Systems*, volume 1236 of *Lecture Notes in Computer Science*, pages 156–170. Springer, 1996.

[3] C. Bremond. *Logique du Recit*. Editions du Seuil, Paris, 1973.

[4] R. A. Brooks. A robust layered control system for a mobile robot. Technical report, Cambridge, MA, USA, 1985.

[5] R. A. Brooks. Intelligence without representation. pages 343–362, 1995.

[6] M. Cavazza, R. Santos de la Camara, M. Turunen, and the Companions consortium. How was your day ? a companion eca. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, accepted for publication (demonstration paper)*, Toronto.

[7] M. Cavazza, C. Smith, D. Charlton, L. Zhang, M. Turunen, and J. Hakulinen. A ŚcompanionŠ eca with planning and activity modelling (short paper). In M. Padgham, Parkes and Parsons, editors, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent*

Systems (AAMAS 2008), pages 1281–1284, Estoril, Portugal, May, 12-16. 2008.

[8] K. Komatani and A. I. Rudnicky. Predicting barge-in utterance errors by using implicitly-supervised asr accuracy and barge-in rate per user. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 89–92, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[9] K. Matsuyama, K. Komatani, T. Ogata, and H. G. Okuno. Enabling a user to specify an item at any time during system enumeration Ű item identification for barge-in-able conversational dialogue systems. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 252–255, Brighton UK, 6-10 September 2009.

[10] A. Raux and M. Eskenazi. A multi-layer architecture for semi-synchronous event-driven dialogue management. In *ASRU*, pages 514–519, Kyoto, Japan, 2007.

[11] R. C. Rose and H. K. Kim. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2003.

[12] A. R. Setlur and R. A. Sukkar. Recognition-based word counting for reliable barge-in and early endpoint detection in continuous speech recognition. In *Proceeding of the International Conference on Spoken Language Processing*, pages 2135–2138, 1998.

[13] C. Smith, M. Cavazza, D. Charlton, L. Zhang, M. Turunen, and J. Hakulinen. Integrating planning and dialogue in a lifestyle agent (short paper). In H. Prendinger, J. Lester, and M. Ishizuka, editors, *Proc. of the 8th Int. Conf. on Intelligent Virtual Agents (IVA 2008)*, page 146Ű153. LNAI 5208, 2008.

[14] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 766–773, New York, NY, USA, 2002. ACM.